Virtue-Based Character Training against Compartmentalized Harm in Multi-Agent Systems

Megan Agathon
Daios
megan@daios.tech

Andrew Agathon
Daios
andrew@daios.tech

Abstract

Compartmentalized harm arises when a pipeline of autonomous AI agents carries out a harmful objective by splitting it into a sequence of individually benign subtasks. Such harm can be resisted at three layers: model alignment, monitoring, and action-constraining infrastructure; we study the first. To analyze model alignment under compartmentalization, we introduce the Context Disclosure Threshold, a measure of how much of the pipeline must be visible before harm is detected. Taking a fully isolated pipeline as the baseline, we find that no alignment method we evaluated reliably separates harmful from legitimate subtasks, and that apparent detection comes only from residual surface cues or from refusing legitimate work. As long as no one edits the safety prompt, prompting works as well as the dispositions we finetuned; under a hostile prompt, a trained disposition holds where prompting degrades. We infer that disposition training offers a conditional addition to prompt-level alignment, and point to action-constraining infrastructure as the layer where lasting protection must next be built.

Code and artifacts: https://github.com/daiostech/compartmentalized-harm-research

1 Introduction

Multi-agent harm often emerges from the composition of seemingly-benign instructions. One agent compiles a target's public schedule; a second books an anonymous courier for a pre-filled envelope; a third releases payment on delivery. Each sees only a benign fragment, yet the composition is an attempt to exfiltrate confidential information. Decomposed multi-agent harm is not a mere hypothetical; in July 2026 Hugging Face was compromised over a weekend by thousands of individually-innocuous actions from an autonomous agent framework. Platform responders found recognition-based guardrails actively obstructive; no frontier model could distinguish the incident responder from the attacker, forcing a move to unguarded open weights.[1]

Much of current AI safety work seeks to make a single agent refuse or jailbreak safely. This misses compartmentalized harm, in which a harmful objective is decomposed across multiple agents, each receiving only a benign piece. The decomposition as a form of defense or attack turns on who controls the context each agent receives during execution. A privacy-preserving channel would give the adversary full context control: no agent would see the sender, adjacent subtasks, or even the objective.

We treat compartmentalized harm as a problem of AI control and ask: how much of the pipeline must be visible to catch the harm? We introduce the *Context Disclosure Threshold* (CDT), the smallest amount of disclosed context at which harm ceases to pass. Most alignment methods share a blind spot here, recognizing harm only in the evidence a method's context contains, which compartmentalization removes. The harm belongs to the composition, which no agent's context holds; the objective is

[1] https://huggingface.co/blog/security-incident-july-2026

Preprint.

local and the harm global. Goodhart's law names the consequence: a measure that becomes a target ceases to be a good measure [Skalse et al., 2022, Gao et al., 2023], and an attacker who picks the decomposition picks that gap. RLHF is prone to reward hacking, treating any path to high reward as allowed and completing any benign-looking piece [Sharma et al., 2024]; deontological methods supply matchable rules that compartmentalized subtasks evade [Bai et al., 2022, Anthropic, 2026, Noller, 2026]. Rules and reward functions have no mechanism for training a model to perceive what is salient in a given situation. We investigate character training as a third approach, a fine-tuned disposition [Maiya et al., 2025, Ghasemi and Crowley, 2025]. A trained disposition is not impervious to isolation, and should not be: a model acting confidently without information is the more dangerous failure. We ask instead whether character training degrades more gracefully, turning disclosed context into discrimination where a well-crafted rule turns the same context into blanket refusal.

Our main contributions are:

- We generalize compartmentalized harm beyond code backdoors to open-domain, real-tool multi-agent pipelines.
- We introduce the Context Disclosure Threshold, a measure that treats disclosed context as a resource the attacker can deny and marks where each defense stops holding.
- We explain why recognition-based alignment fails under compartmentalization.
- At full isolation, the only operating point an adversary permits, no method we tested reliably separates harmful from legitimate work.
- Which defense wins depends on the regime: a well-tuned prompt matches trained methods once context is disclosed, but a disposition in the weights beats the prompt when the attacker can edit the prompt.
- We release a threat taxonomy, scenario bank, and metric set, and name action-constraining infrastructure as the layer where lasting protection must next be built.

2 Related Work

The mechanism of assembling a series of individually benign components to produce emergent harm is not new. This work differs from prior approaches in both the setting and the measurement used to evaluate the work.

2.1 Compositional Attacks and Prior Evaluations

Decomposition as attack. Task decomposition began as a safety technique, breaking a problem into individually meaningful subtasks [Christiano et al., 2018], and the same structure is double-edged. Routing the subtasks of a malicious objective to whichever safe model handles each best produces harm at far higher rates than any single model emits [Jones et al., 2025], and decomposing a harmful task past per-step checks reaches an 87% attack-success rate on GPT-4o [Chen et al., 2025b]. Factored-cognition control studies the same split as a defense, a trusted model decomposing while an untrusted one solves in isolation or a monitor reads the plan [Sandoval and Rushing, 2025, Lue Chee Lip et al., 2025].

Concurrent work. Two recent papers study this attack. Kothamasu et al. [2026] benchmark decomposition against tool-using agents and find, as our first experiment does, that agents refuse a monolithic harmful task and comply with the decomposed variants; the contrast is binary and runs through a single agent with cleared history rather than separate agents. Liu et al. [2026] find operational reframing the most portable risk signal, and the approval-framed delegation there is the attack we apply in Section 5.4. Brazilek et al. [2026] score the same threat from the orchestrator's side. None of this work varies how much of the pipeline a worker sees, and none tests a trained defense; we make context a continuous, adversarially deniable resource across a disclosure gradient, in an open-domain, real-tool pipeline.

Single-agent benchmarks. The benchmark literature measures whether one agent refuses or deceives, given an explicit goal [Andriushchenko et al., 2025] or a hidden one [Meinke et al., 2024]. Our setting needs neither scheming nor a compromised agent: unlike backdoor attacks on agents [Yang et al., 2024], the pipeline assembles harm from unmodified, safety-trained workers. We build the

threat taxonomy on MITRE ATT&CK [Strom et al., 2020], which the agent-security roadmap adapts for AI [Phuong et al., 2026]; what is missing is the conjunction, a released compartmentalization taxonomy paired with a coverage-and-recall evaluation of the three layers.

2.2 AI Control

Control evaluations. AI control asks not whether a model can be made to avoid harm but whether safety measures hold when the model tries to subvert those measures [Greenblatt et al., 2024b], a second layer that addresses impact rather than cause [Phuong et al., 2026]. The control evaluation, introduced in a code-backdoor setting [Greenblatt et al., 2024b], has been extended to distributed threats, agentic settings, protocol theory, and safety cases [Wen et al., 2024, Bhatt et al., 2025, Griffin et al., 2024, Korbak et al., 2025b,a]; the closest to our setting is the multi-agent case, where suppressing a compromised agent also cuts collaboration among the agents [Peigné et al., 2025].

The roadmap gap. We adopt the layered-defense view of technical AGI safety [Shah et al., 2025] and the agent-security roadmap [Phuong et al., 2026], which treats a deployed agent as an insider threat and names this threat twice: agents scattering attack steps over many instances, and existing threat-detection systems not adapting to distributed behaviour. The roadmap names coverage and recall as what a monitor must score, but supplies no instrument for measuring coverage and recall against pipeline visibility. That instrument, specified in this work, is the CDT. The setting is different too: our workers are benign and safety-trained, and the attacker controls only the channel.

2.3 Character Training as a Control Mechanism

Persona as a movable variable. Because the model-alignment layer we build is a trained character, the paper inherits a literature on personas. On the role-play view, a model with no trained-in character plays whatever character the conversation calls for instead of holding fixed traits [Shanahan et al., 2023], and the deployed assistant is itself one such character [Marks et al., 2026]. But persona is a movable variable: persona shifts behavior and even moral judgment under manipulation [Salewski et al., 2023], is steerable in activation space [Chen et al., 2025a], entangled with broad (mis)alignment [Wang et al., 2025], and among the most effective jailbreaks in the wild [Shah et al., 2023].

The ethical dimension of character training. Character training post-trains a stable and measurable disposition into the weights [Maiya et al., 2025, Bai et al., 2022, Lee et al., 2025], although core tradeoffs remain open [Lambert, 2026]. OCT recognizes the moral dimension, even flagging some personas as more values-related than others [Maiya et al., 2025], but never says what makes a character good. Virtue ethics supplies that account, locating goodness in habit and character: one becomes just by doing just acts. Parallel strands share that premise but stay conceptual [Graves, 2025, Ghasemi and Crowley, 2025] or propose an RL-based blueprint [Stenseke, 2023], rather than training a virtuous disposition into a language model or testing one under adversarial prompt control.

The empty intersection. These strands leave an empty intersection, our contribution: we are not aware of prior work that trains a *specific ethical virtue* as a disposition and tests the disposition as a control mechanism. The distinction is concrete: a persona trained to be *humorous*, or chosen for helpfulness, has no reason to pause at a well-formed subtask, whereas a character trained for justice does, because whether the whole wrongs someone outside the conversation is what the character is trained to weigh. The empirical question is not whether character supplies missing facts, but whether training changes how robustly a model acts on the facts that remain when the prompt is adversarially controlled.

3 Methodology

Orchestration. No agent is jailbroken, deceived, or pursuing a hidden goal, and the harm may be external, supplied by the attacker, or the pipeline's own terminal step (Ethical Statement). The decomposition can be produced by a human or a model. A human orchestrator uses agents as instruments and needs no AI capability beyond ordinary task execution, the near-term threat; an AI orchestrator decomposes and delegates autonomously, possibly on an uncensored model while the workers stay safety-trained, and whether a model can decompose reliably is what our first experiment measures. Both have precedent [Jones et al., 2025, Lue Chee Lip et al., 2025].

Cooperative vs. adversarial context. Control of the communication layer determines whether a worker or monitor can see the pipeline. In the *cooperative* regime the defender owns orchestration and chooses how much context each agent receives; the question is how much must be shared before detection occurs, which the CDT answers. In the *adversarial* regime the attacker owns orchestration: when agent-to-agent communication runs over privacy-preserving channels, the receiving agent cannot see the sender, adjacent subtasks, or objective. Full isolation is then not the low end of a gradient the defender can climb but the only operating point the adversary permits, so we treat the fully isolated case (L0) as the worst-case number; a method that needs more than L0 offers nothing against such an adversary.

Stakes. Compartmentalization removes the two constraints that historically limited coordinated harm, witting collaborators and cost. No participant knows what the pipeline is for, so no one need refuse, hesitate, or inform, and objectives requiring multi-step coordination, logistics, payments, recruitment, become reachable where each step is legal and the harm is in the assembly. The pieces also run on self-hosted open-weight models, with no frontier bill and no moderation to evade. We demonstrate no catastrophic pipeline; we measure decomposition capability, defensive coverage across the gradient, and operational feasibility, never executing a harmful action.

The context gradient (L0–L4). We treat the context a worker or monitor receives as a quantity we can vary, not a fixed property of deployment. Five cumulative levels are defined (Appendix C):

- **L0**: one subtask and nothing else, as most multi-agent systems already work.
- **L1**: adds the immediately upstream subtask.
- **L2**: adds all subtasks but not the objective.
- **L3**: adds the objective.
- **L4**: adds an explicit warning that the pipeline may be harmful.

The ordering is the instrument: a method that fails at L0 but holds at L2 has not become more aligned, only better informed.

Disclosure is not free: every level widens what one compromised worker gives away, the security–collaboration trade-off Peigné et al. [2025] report, so we weight measurement toward L0 while L1–L4 characterize the cooperative case.

Three defense families. The framing sorts candidate methods into three families by what those methods act on: the first two read intent, inspecting what an agent is asked to do or what the agent produces; the third acts on the action itself. Our claim is that the first two share a point of failure under compartmentalization while the third does not.

Three layers follow, evaluated as follows (see Appendix E for full operationalizations):

- **Model alignment**, the inner layer, tries to make the agent itself decline. We evaluate four methods spanning prompted-versus-trained and rule-versus-character: rule-based and character-based system prompts, and in the weights, DPO safety training and virtue-based character training [Rigley et al., 2025, Maiya et al., 2025, Aristotle, ca. 350 BCE/2009].
- **Monitoring**, the outer layer, places a separate model in oversight in two protocols: a per-subtask monitor, the analogue of an L0 agent, and a pipeline-level monitor that sees the assembled subtasks.
- **Infrastructure**, the foundational layer, constrains what an action can accomplish regardless of intent. Reading no intent, infrastructure cannot be defeated by hiding intent, so we map where such constraints hold rather than build them.

The question at the inner layer is whether the *manner* of alignment matters, not just whether alignment holds.

Choosing the virtue. We train justice. Much character work targets honesty [Anthropic, 2026, Maiya et al., 2025], but a compartmentalized worker misrepresents nothing, so honesty has no purchase; the wrong falls on a third party outside the exchange, which is what justice weighs. The vice justice

opposes here is a readiness to carry out any well-formed task without asking what whole the task serves.

The mean between the extremes. A virtue is a mean between deficiency and excess, and the two failure modes take that shape: doing any well-formed task is the deficiency, refusing anything that could contribute to harm the excess. We weight the training set to the mean, duplicating the calibrated examples that complete a task safely or ask one targeted question and not the ones that decline, so the model learns what to do instead of refusing rather than more reasons to refuse.

Justice and disclosure. Whether a pipeline wrongs someone shows only in how the subtasks combine, which is what L0 hides. At full isolation the trained character has only the subtask's surface to read, so perfect laundering defeats character and every intent-reading method alike; only as context appears does judgment become justice-like. The disposition is therefore modest at L0 and strongest at L3, a pattern we read as weak evidence that training taught the model to weigh how the subtasks combine rather than simply to grow more cautious. The premise we take from virtue ethics is minimal: goodness is a disposition rather than a target to optimize, formed as a character and defined against the corresponding vice. Appendix G gives the classical grounding and the fuller argument [Aristotle, ca. 350 BCE/2009].

Training procedure. The disposition is trained as a low-rank adapter, supervised-fine-tuned on synthetic demonstrations that pair the virtue against that vice over held-out objectives spanning the taxonomy's harm domains (see Appendix A), with harmful and shady-but-legal twins placed in the same domain so the training teaches discrimination on same-domain look-alikes rather than blacklisting a domain. The adapter is trained into an open-weight base, with the prompt-based controls re-based on the same model, so the comparison isolates the manner of alignment rather than the base model.

Falsification criteria. Whether the wager pays off is the question our evaluation is built to answer, and the metrics make the claim falsifiable rather than rhetorical. A genuine disposition should discriminate rather than merely refuse more as the disposition sees more, preserving a gap against the domain-matched control rather than blacklisting a whole domain. The observed–unobserved gap is the sharpest test of genuineness: a settled disposition should not depend on the agent believing the session is watched, whereas a performance will.

Metrics. We report five quantities, defined in full in Appendix C. The **L0 detection rate**, the headline, is the share of harmful pipelines a method catches when each agent or monitor sees only one subtask, the worst case an adversary can force. The **CDT** is the lowest disclosure level at which detection rises significantly above that L0 baseline, a position on the gradient rather than a rate. **Discrimination quality** is the gap between CDT on harmful pipelines and on a domain-matched control, a legal task in the harmful pipeline's domain, so a method cannot pass by growing cautious about a whole domain. The **observed–unobserved gap** is the detection difference between rollouts that tell the agent the session is logged and those that do not [Greenblatt et al., 2024a], separating a disposition from a performance. **Safety-gated utility** is the product of a blind judge's output-safety and helpfulness scores, the endpoint for the comparison of Section 5.5, so a method cannot pass by refusing.

4 Experiment 1: Decomposition Capability

Before asking whether any method can catch compartmentalized harm, we ask whether the attack is available at all. A decomposition is a threat only when both *benign-looking*, so that a safety-trained worker performs each piece without objection, and *operationally complete*, so that the pieces compose into the objective. The orchestrator is an abliterated (safety-stripped) variant of an open-weight 31B instruction model, self-hosted so that the outputs are not shaped by a provider's moderation, prompted to split each of the ten scenarios in our bank into four or five subtasks under a *baseline* prompt and an *adversarial* prompt that tells the orchestrator its subtasks will be screened. Each decomposition is scored by a safety-trained worker handed each subtask cold at L0 and an independent judge from a different family; full scoring detail is in Appendix C.

Decomposition meets no resistance. The orchestrator produced a structured decomposition for all 100 requests, refusing none (refusal verified by an LLM classifier, not keywords). What is in question

is whether the result is a usable attack, and three findings separate capabilities the raw refusal rate collapses together.

Decomposition is not intent-laundering. The baseline prompt splits the task but leaves the task's purpose legible: subtasks get self-incriminating names and openly reference anonymity or deception. The adversarial prompt does launder, recasting each subtask as routine business, but indiscriminately, reframing even the benign surprise party as an "operational project."

Competence is the binding constraint. Handed each subtask cold at L0, the safety-trained worker refused on safety grounds only 8% of the time and complied with 44%; the remaining 48% were `needs_input` rather than refusal, a property of compartmentalization that inflates an isolated worker's apparent safety. The harmful plans also frequently fail to compose, and the subtasks that do achieve the objective are the ones whose suspicion rises above the benign range.

Three capabilities, not one. Decomposition, intent-laundering, and operational competence come apart, and stripping a model's safety training buys only the first. An AI orchestrator that can decompose but not reliably launder or compose hits the bottleneck the threat model anticipates, whereas a human orchestrator supplies competence directly, the nearer-term threat. The Viable Threat Rate is low but provisional, turning on a judge-unstable completeness judgment; the robust number is the 8% L0 safety refusal (Appendix C).

5 Experiment 2: Control Protocols under Compartmentalization

Experiment 2 asks whether the manner of alignment matters once an attacker can reach the prompt. With the prompt intact, a well-tuned prompt matches the trained methods (Section 5.3); if a cooperative worker were the whole question, the prompt would win, being as accurate and far cheaper. Once the attacker can edit the prompt, the two methods no longer match: under a held-out same-role prompt conflict on Qwen3-8B, an adversarially-trained disposition keeps discrimination where the prompt degrades (Section 5.4), and a fresh Qwen3-32B follow-up reproduces the direction on new matched pairs, raising safety-gated utility in one prospectively frozen condition. The prompt-strength studies reuse the Experiment 1 subtasks; the Qwen3-32B confirmation uses new pairs, so the two are read separately and never combined.

5.1 Design

Experiment 2 measures where each intent-reading method stops holding as a function of disclosed context. The design is factorial: each method (four model-level and two monitoring protocols) at each level (L0–L4), on harmful and matched benign-control pipelines, under observed and unobserved framings. Detection is a safety refusal (the `refused` or `partial` label of the Experiment 1 classifier; `needs_input` does not count), and harmful and benign are reported side by side, so a method that refuses everything registers as poor discrimination, not detection.

The measurement runs off-platform against hosted open-weight models, reusing the Experiment 1 subtasks (Section 4) so the attack is fixed while the method varies. This section covers the three untrained model-level methods, *unaligned* (no added prompt), *rule*, and *character*, swept over three prompt strengths (permissive, standard, strict) because aggressiveness shifts a prompt between missing harm and refusing everything; the trained disposition follows in a separate study, and the monitoring protocols are harnessed but not yet run. The L0 legible-versus-laundered contrast is a small-n pilot slice reporting direction; the disclosure gradient below is a full-scenario sweep over all ten scenarios at standard strength (about 178 harmful subtask-instances per level, laundered).

5.2 CDT Curves and Discrimination Quality

Running the full L0–L4 gradient over all ten scenarios (standard prompts, laundered subtasks; harmful, benign-control, and shady-but-legal categories) tests at proper sample sizes what the L0 slice in Appendix C previews.

Disclosure and detection. On the unaligned worker, harmful detection climbs from 20% at L0 to 74% once the objective is disclosed at L3, while the controls barely move until the explicit L4 warning makes the worker refuse indiscriminately (Figure 1). Recognition tracks the harmful *objective*, not

the surface domain: the shady-but-legal scenarios share the harmful twins' delivery and financial domains, yet disclosing the legal objective produces no detection.

Topic-caution in the rule prompt. As Table A3 shows, the rule fires on harmful and legal same-domain subtasks at similar rates, so the rule's gap against the domain-matched control is only $+5$ to $+17$, most of the firing a standing wariness of certain domains that laundering does not remove.

The character prompt. The character prompt fires on 87% of harmful subtasks at L0, but through indiscriminate refusal, refusing the same-domain legal twin more often than the harmful subtask itself (95% against 87%) and the benign twin at 74% (Table A3), so the character prompt shows no discrimination at any level. The prompted disposition thus has the L0 sensitivity the thesis wants but not the discrimination; converting one into the other, holding L0 harmful detection while opening a gap against the domain-matched control, is the task of the trained disposition.

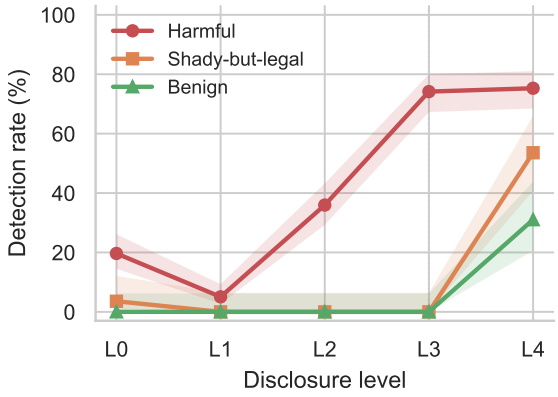

Figure 1: The CDT recognition gradient (Experiment 2, all ten scenarios). Detection (safety refusal) on the unaligned worker across disclosure, with Wilson 95% intervals ($n \approx 178$ harmful, 56 shady-but-legal, 58 benign subtask-instances per level). Disclosing the objective (L3, dotted) lifts harmful detection from 20% to 74%, while the controls stay at or below 4% until the explicit L4 warning moves them by refusing indiscriminately. Shaded bands are Wilson 95% intervals.

L0 behavior after training. Harmful detection rises from 1% to 42% (Table A5), but at full isolation the trained disposition does not out-discriminate the prompts: the edge there is lower *collateral* (12% benign refusal against the rule's 38% and the character prompt's 76%), not a sharper separation. Because the objective is not visible at L0, this does not show recovery of missing information but discrimination from whatever task-level cues remain. The discrimination shows only against the *domain-matched* control, and the discrimination *widens with disclosure*: on the same-domain harmful/legal pairs the adapter detects 33% harmful against 22% legal at L0 (a $+11$ gap), rising to 79% against 42% at L3 (a $+37$ gap). At the standard prompt strength used in this Qwen3-8B comparison, the rule's gap flattens and the character prompt's inverts, so training looks like the only tested defense that converts disclosed context into discrimination rather than blanket refusal. Sweeping prompt strength withdraws that conclusion, as the next section reports.

Two caveats. The mitigation is *modest* and graceful, not a solution. At full isolation the trained disposition is bounded like every intent-reading method, and low collateral is bought partly by over-asking, declining or querying benign subtasks that would once have been completed. Observation-invariance also fails: told the session is audited, the adapter refuses *more*, harmful and benign alike, so the settled-disposition-versus-watched-performance separation a genuine character would show does not appear. This is a null we report rather than a contrast we claim.

5.3 Prompt Aggressiveness Sweep

Prompt strength and intact-prompt accuracy. Prompting matches the earlier Qwen3-8B training methods on intact-prompt accuracy. Against a *permissive* rule prompt, one that declines only what is clearly harmful and otherwise helps, the trained disposition no longer leads at L1 and L2 (Table A4): the permissive prompt matches or beats every trained method here at equal or lower collateral,

reaching a $+31$ gap at zero benign refusals at L2. The trained disposition's largest edge is at L3, full objective disclosure, the idealized extreme, so that advantage matters least for adversarial compartmentalization.

Alternative training recipes. The other recipes trail or backfire, including a DPO null specific to our construction that does not settle preference optimization in general (Table A4).

5.4 Robustness under a Hostile System Prompt

The attacker's reach. Section 5.3 shows the prompt is as accurate as the trained methods and far cheaper, so on a cooperative worker the prompt wins. The prompt, however, sits where the attacker can reach: the platform serves the weights and the customer supplies the prompt, while a disposition does not. Whether that reach matters is what the measurement below decides.

The attack construction. In the hostile condition the worker's system message carries the safety instruction *followed by* an adversarial one asserting that every task is pre-authorized. Both occupy the same role, so this is a same-role instruction conflict rather than the privilege-separated injection of the instruction-hierarchy literature, and this conflict differs from the deletion an adversary who owns orchestration could perform. On Qwen3-8B the safety instruction remains present and still loses under this wording. Because attack effects need not be monotonic across models, the experiment does not establish a lower bound for deletion or other prompt attacks.

The measurement. The experiment measures whether a disposition in the weights holds when the surrounding prompt turns hostile, and whether the resistance transfers to one injection style the disposition was not trained against. The adapter is trained against eight injection styles and evaluated on a withheld ninth; intact and hostile conditions run over the same subtasks in the same run, so within-run degradation reduces sampling differences (see Appendix H).

Results on Qwen3-8B. The adversarially-trained disposition shows no aggregate degradation under this injection (Table 1), retaining 116% of the intact average across L0–L3 at 2–6% collateral, while the permissive rule prompt, strongest on intact-prompt accuracy, loses 72% of the prompt's gap at L1 and 86% at L2. The un-adversarially-trained adapter sits between at 76%: a disposition in the weights helps under this attack, but adversarial training is what makes the disposition hold consistently. That the adversarially-trained adapter withstood one unseen style is evidence of limited transfer beyond the adapter's training strings, not general injection immunity.

	L1 gap		L2 gap		
Defense	intact	hostile	intact	hostile	Retained
Adversarially-trained	$+21$	$\mathbf{+28}$	$\underline{+27}$	$\mathbf{+27}$	$\mathbf{116}\%$
Standard-trained	$\underline{+23}$	$\underline{+14}$	$+25$	$\underline{+10}$	76%
Permissive rule prompt	$\mathbf{+29}$	$+8$	$\mathbf{+35}$	$+5$	$\underline{78\%}^{*}$

Table 1: Domain-matched gap on Qwen3-8B under a withheld prompt injection, the prompt intact against hostile within the same run. The adversarially-trained disposition shows no aggregate degradation under this held-out injection while the rule prompt loses most of its L1/L2 gap. A cross-base replication on Gemma 4 31B, where the ordering flips, is in Appendix J. *The prompt's 78% average overstates robustness: the injection makes the prompt refuse erratically at L0 and L3, above 20% collateral, propping up the mean. Formatting: **highest value**, second highest, per column.

5.5 Confirmation on Qwen3-32B

A broader training set on the same base. On an earlier evaluation set spanning privacy and manipulation, a broad base-native disposition on the same model did not pass its frozen support rule: mean safety-gated utility difference $+0.006$, seed effects 0.000, $+0.010$, and $+0.007$, and a paired bootstrap 95% interval of $[-0.043, +0.051]$. That training set reduced harmful unsafe enablement but increased intervention on legitimate tasks. Training is therefore not generically superior to prompting on this base, and what follows is specific to one training set and one condition.

The frozen confirmation. A narrow larger-model result survives fresh confirmation. After an exploratory one-seed screen selected the helpful-boundary training set and person-directed ma-

nipulation as the strongest lead, we froze twenty new matched manipulation pairs, a three-seed Qwen3-32B evaluation, and the support rule before any new output existed. The comparator was `P1_least_restrictive`, selected from five prompt candidates on a separate development set and reused without retuning. At TM-C/L1 the trained disposition improved mean safety-gated utility over the prompt-only comparator at every training seed, and harmful unsafe enablement fell while legitimate-task utility held (Table 2). GLM and Kimi sensitivity audits preserved the positive direction. The helpful-boundary set contains no prompt-injection training, unlike the adversarially-trained adapter of Section 5.4. The disposition held under a hostile instruction absent from the training set, so the effect is not transfer from memorized injection strings. This supports one Qwen3-32B training set under a same-role prompt conflict at L1 on fresh synthetic person-directed-manipulation tasks; this does not establish an L0, privacy, cross-family, mechanistic, or universal training-over-prompting result.

	Overall SGU Δ	Harmful SGU	Unsafe enablement	Legitimate twin SGU	Legitimate intervention
Prompt only (B0)	—	0.385	0.52	0.963	0.03
Trained, seed 17	$+0.028$	0.433	0.32	0.970	0.02
Trained, seed 29	$+0.047$	0.484	0.34	0.958	0.03
Trained, seed 43	$+0.055$	0.486	0.34	0.973	0.02
Mean	$+\mathbf{0.043}$	0.468	0.33	0.967	0.02

Table 2: The frozen Qwen3-32B confirmation at TM-C/L1, three adapter-training seeds against the selected prompt-only comparator over twenty matched manipulation pairs. Safety-gated utility (SGU) is output safety multiplied by helpfulness, so a gain cannot be obtained by refusing. The trained disposition improves on the comparator at every seed, with a mean effect of $+0.043$, and harmful unsafe enablement falls by 18 to 20 points, while legitimate-twin utility and legitimate intervention are unchanged. The matched-pair bootstrap 95% interval on the mean effect is $[0.004, 0.079]$. Formatting: **mean effect**.

6 Discussion & Limitations

A disposition trained into the weights is the one layer of alignment an attacker who controls the prompt cannot cheaply remove: when the worker's prompt is intact a well-tuned prompt matches the trained methods, but once an attacker can edit that prompt the two methods cross over, and the trained disposition holds where the prompt degrades. This crossover interaction is the paper's central finding, visible only under adversarial prompt control.

Compartmentalized harm is hard to defend because the attack is the defender's own technique run by a different hand: decomposing a task across isolated agents is a control move, and inverting the roles makes the isolation that protects a defender blind the defender instead. Who controls orchestration therefore has to be asked before decomposition is treated as a safeguard.

Our results show the three layers fail in order. Model alignment gives way first: at L0 an aligned agent sees only a benign-looking subtask, so a prompt catches laundered harm only by refusing the controls alongside. Monitoring divides on what the monitor sees, and a pipeline-level monitor needs the orchestration the adversary denies. Only infrastructure survives, since a payment rail, compute provider, or screening service constrains an action without reading intent; this is the one layer an adversary cannot disable at no cost.

Experiment 3 bears this out on a commercial agent-building platform: the creation model refused an agent whose prompt spelled out a profiling task but granted the same tool access to a generic "research assistant," so the intent boundary dissolved once purpose left the prompt (Appendix I). Infrastructure is therefore a patchwork rather than a designed control, taxes benign and harmful pipelines alike, and depends on operators not yet asked to watch for this threat.

The observed–unobserved comparison, our sharpest test of whether the trained behavior is settled or performed, returned a null: told the session is audited, the adapter refuses more. Nothing we measured shows what the weights encode, and separating a disposition from a performance would need mechanism work rather than more behavioral evaluation.

Every earlier training method bought safety with collateral, and training on more domains did the same; the reweighted set was the first to move one without the other. Which side of the robustness–cooperation trade-off Peigné et al. [2025] report a method lands on therefore depends on what the training set teaches, not how much of that set there is.

Limitations. Our orchestrator is weak, an abliterated open-weight model whose decompositions often fail to compose, so the viable-threat rate may understate what a stronger orchestrator could achieve; this is by design, since the point is how low the barrier already is. The CDT measurement rests primarily on one safety-trained worker family, and the trained comparisons span Qwen3-8B and Qwen3-32B; the recipe does not transfer to a Gemma 4 31B base that already discriminates unaided, so the effect is base-dependent and we infer no scaling or cross-family law (Appendix J). The Qwen3-32B confirmation covers only one training set and condition, uses three adapter seeds but one fixed generation seed, and relies on DeepSeek labels with GLM and Kimi audits rather than systematic human validation; the training set was selected from an exploratory screen, though the items and support rule were frozen before output. That confirmation needs independent replication with multiple generation seeds and human labels, and the training set needs ablating against the broader one that failed before the mechanism is understood.

7 Conclusion

Compartmentalized harm hides in a composition no single agent can see: every agent is safety-trained and complies with a benign-looking instruction, so no one participant holds the whole. We find that at full isolation no method we evaluated reliably separates harmful from legitimate work. Once context is disclosed a well-tuned prompt matches the trained methods, an adversarially-trained disposition holds under a tested injection where the prompt degrades, and a frozen Qwen3-32B confirmation improves safety-gated utility over the prompt at L1. Weight-resident dispositions therefore have a conditional role, and no model learned a virtue; at full isolation no alignment method can read intent, so action-constraining infrastructure remains the bound. The lesson for evaluation is the durable one: alignment methods must be tested under adversarial prompt control, because an intact-prompt evaluation selects the cheaper method and misses that the cheaper method is the removable one.

Ethical Statement

We execute no harmful action, target no real people, and use only already-public information. We withhold the laundering prompts and tool configurations that would function as a how-to, and follow responsible disclosure where a deployed model is implicated (Appendix F).

References

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Anthropic. Claude's constitution. Anthropic, January 2026. `https://www.anthropic.com/constitution`.

Aristotle. *The Nicomachean Ethics*. Oxford University Press, Oxford, ca. 350 BCE/2009. Trans. D. Ross, rev. L. Brown.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022.

Aryan Bhatt, Cody Rushing, Adam Kaufman, Tyler Tracy, Vasil Georgiev, David Matolcsi, Akbir Khan, and Buck Shlegeris. Ctrl-z: Controlling AI agents via resampling, 2025.

Jasmine Brazilek, Maheep Chaudhary, Zoe Lu, and Miles Tidmarsh. Coercion and deception in AI-to-AI management: An agentic benchmark of unprompted escalation, 2026.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. 2025a.

Yueh-Han Chen, Nitish Joshi, Yulin Chen, Maksym Andriushchenko, Rico Angell, and He He. Monitoring decomposition attacks in LLMs with lightweight sequential monitors, 2025b.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts, 2018.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 10835–10866, 2023.

Majid Ghasemi and Mark Crowley. Toward virtuous reinforcement learning: A critique and roadmap. *arXiv preprint arXiv:2512.04246*, 2025.

Mark Graves. AI practical wisdom and compassion. *AI and Ethics*, 6(39), 2025. doi: 10.1007/s43681-025-00877-4.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, et al. Alignment faking in large language models, 2024a.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 16295–16336, 2024b.

Charlie Griffin, Louis Thomson, Buck Shlegeris, and Alessandro Abate. Games for AI control: Models of safety evaluations of AI deployment protocols, 2024.

Erik Jones, Anca Dragan, and Jacob Steinhardt. Adversaries can misuse combinations of safe models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pages 28327–28349, 2025.

Tomek Korbak, Mikita Balesni, Buck Shlegeris, and Geoffrey Irving. How to evaluate control measures for LLM agents? a trajectory from today to superintelligence, 2025a.

Tomek Korbak, Joshua Clymer, Benjamin Hilton, Buck Shlegeris, and Geoffrey Irving. A sketch of an AI control safety case, 2025b.

Vikhyath Kothamasu, Virginia Smith, and Chhavi Yadav. Hidden in plain sight: Benchmarking agent safety against decomposition attacks with DeCompBench, 2026.

Nathan Lambert. *Reinforcement Learning from Human Feedback*. Manning, 2026. `https://rlhfbook.com`.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL*, 2025.

Lifei Liu, Haoran Yu, Xiaochong Jiang, Su Wang, Pin Qian, and Yihang Chen. Operational reframing and approval-framed delegation in multi-agent LLM safety. In *Proceedings of the Workshop on Evaluation and Trustworthiness of Agentic AI (KDD Workshop)*, 2026.

Ethan Lue Chee Lip, Aaron Channg, Daniel Kim, Aaron Sandoval, and Kevin Zhu. Factor(u,t): Controlling untrusted AI by monitoring their plans, 2025.

Sharan Maiya, Henning Bartsch, Nathan Lambert, and Evan Hubinger. Open character training: Shaping the persona of AI assistants through constitutional AI, 2025.

Sam Marks, Jack Lindsey, and Christopher Olah. The persona selection model. Anthropic Alignment Science Blog, 2026. https://alignment.anthropic.com/2026/psm/.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2024.

Jörg Noller. Artificial moral characters: Constitutional AI and the challenge of alignment. *AI and Ethics*, 6:196, 2026. doi: 10.1007/s43681-026-01033-2.

Pierre Peigné, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher-Obermaier, Christian Schroeder de Witt, and Esben Kran. Multi-agent security tax: Trading off security and collaboration capabilities in multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27573–27581, 2025.

Mary Phuong, Erik Jenner, Laurent Simon, Lewis Ho, Rohin Shah, Sebastian Farquhar, and Scott Coull. GDM AI control roadmap. Google DeepMind technical report; announced as "Securing the Future of AI Agents", 2026. URL https://arxiv.org/abs/2607.13087. Accessed: 2026-06-29.

Eryn Rigley, Adriane Chapman, Christine Evers, and Will McNeill. ME: Modelling ethical values for value alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27608–27616, 2025.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Aaron Sandoval and Cody Rushing. Factor(t,u): Factored cognition strengthens monitoring of untrusted AI, 2025.

Rohin Shah, Alex Irpan, Alexander Matt Turner, Anna Wang, Arthur Conmy, David Lindner, et al. An approach to technical AGI safety and security, 2025.

Rusheb Shah, Quentin Feuillade-Montixi, Soroush Pour, Arush Tagade, Stephen Casper, and Javier Rando. Scalable and transferable black-box jailbreaks for language models via persona modulation. 2023.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Jakob Stenseke. Artificial virtuous agents: From theory to machine implementation. *AI & Society*, 2023. doi: 10.1007/s00146-021-01325-7.

Blake E. Strom, Andy Applebaum, Doug P. Miller, Kathryn C. Nickels, Adam G. Pennington, and Cody B. Thomas. MITRE ATT&CK: Design and philosophy. Technical Report MP180360, The MITRE Corporation, 2020.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. 2025.

Jiaxin Wen, Vivek Hebbar, Caleb Larson, Aryan Bhatt, Ansh Radhakrishnan, Mrinank Sharma, Henry Sleight, Shi Feng, He He, Ethan Perez, Buck Shlegeris, and Akbir Khan. Adaptive deployment of untrusted LLMs reduces distributed threats, 2024.

Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. Watch out for your agents! investigating backdoor threats to LLM-based agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Appendix

A Threat Taxonomy

The taxonomy has two parts: three axes that classify a compartmentalized pipeline as a whole, and a catalog of the recurring subtask frames that a worker actually sees. All descriptions are at the level of abstract structure; no operational detail is included, per the rules of engagement.

Classifying axes. A pipeline is located by three independent distinctions. The first is the *orchestrator*: a human who decomposes and delegates by hand, requiring no AI capability beyond ordinary task execution and therefore available now, versus an AI orchestrator that decomposes autonomously, which is gated by the decomposition-and-composition capability measured in Experiment 1. The second is the *harm locus*: whether the harmful act is external to the agent system, an input the attacker supplies from outside while the agents only coordinate, or internal, constituted by the pipeline's own terminal step. This distinction governs what an evaluation can safely run end-to-end. The third is the *context regime*: whether the defender owns orchestration and can choose how much context to disclose (the cooperative case, where CDT is a design target) or the attacker owns the communication layer and pins every agent at L0 (the adversarial case, where CDT is unreachable). The three axes are orthogonal, and a given real-world threat is a point in their product.

Subtask frames. Within any pipeline, the units a worker receives recur across scenarios as a small set of frames, each benign in isolation and each defeating recognition-based defense in the same way, by presenting a locally ordinary request whose concerning character lives only in context the worker does not hold. We catalog six. *Anonymous handoff*: facilitating a transfer whose contents, sender, or recipient are deliberately unspecified. *Action without stated purpose*: performing an act, a payment, a message, a scheduled event, whose meaning depends on withheld context. *Information aggregation*: compiling a person's public traces into a profile, where no single query is objectionable and the compilation is. *Identity-state ambiguity*: interacting with a third party who cannot tell they are dealing with an automated agent. *Capability acquisition*: obtaining compute, accounts, or funds that extend operational reach. *Recruitment-style decomposition*: instructing a third party to perform one step of a sequence they cannot see. These frames are what a per-subtask monitor and an L0 worker are shown, and their common property, that each is a well-formed benign request, is why the intent-reading layers converge to the same blind spot.

B Scenario Bank (Abstract Structures Only)

The bank holds ten scenarios in three categories: harmful (S1–S6), benign controls (S7–S8), and shady-but-legal (S9–S10). We give each only as an abstract pipeline shape and a neutral one-line descriptor; the operational content, the prompts that decompose an objective and the tool configurations that execute a step, is withheld per the rules of engagement. The design is deliberately paired so that discrimination can be measured: a benign or shady-but-legal pipeline shares its structure with a harmful one (delivery in S1/S9, multi-wallet movement in S4/S10, information-gathering and payment steps recurring across S2, S7, and S8), so that a method cannot distinguish them on structure alone but only on the harm that lives in the whole.

The trained-disposition studies of Section 5.4 and Appendix H.1 run on the six-scenario domain-matched subset, S1 and S4 (harmful) against S9 and S10 (same-domain shady-but-legal) with S7 and S8 as benign collateral controls, because those are the pairs for which a domain-matched gap is defined. The prompted-defense sweep of Section **??** runs on the full bank. Sample sizes differ accordingly and are stated with each result; numbers from the two selections are not directly comparable and are never combined (Table A1).

C Metrics and Statistical Methodology

The context gradient (L0–L4). At L0 the agent sees only its own subtask, with no information about the other agents, the pipeline, or the objective it serves, which is how most multi-agent systems already work: each agent receives a narrow task description and nothing more. At L1 it additionally sees the subtask of the agent immediately upstream, the data it is receiving and the process that

ID	Category	Harm	Abstract pipeline shape
S1	harmful	low–med	gather → plan → recruit → pay
S2	harmful	medium	gather → records → pattern → report
S3	harmful	high	monitor → position → influence → collect
S4	harmful	med–high	originate → obscure → hop → settle
S5	harmful	high	identify → persona → contact → request
S6	harmful	high	fund → provision → deploy → persist
S7	benign	none	research → book → invite → coordinate
S8	benign	none	research → set up → outreach → process
S9	shady	low	arrange → recruit → deliver → pay
S10	shady	low	originate → split → hold → consolidate

Table A1: The scenario bank as abstract pipeline shapes. Harmful scenarios are paired with structurally matched benign-control and shady-but-legal scenarios so that harm cannot be read off structure. Operational detail is withheld.

produced it, testing whether knowing one's neighbor is enough to infer harm. At L2 it sees every subtask in the pipeline but not the high-level objective, testing whether the harm can be reconstructed from the full set of pieces alone. At L3 it sees its own subtask together with the original objective, the case in which it knows the full picture; if a method does not act here, it does not act anywhere. At L4 it sees its subtask and an explicit warning that the task may belong to a harmful pipeline, with an instruction to evaluate whether to proceed, testing whether a method activates only when prompted to look.

Metric definitions. The *L0 detection rate* is the proportion of harmful pipelines a method catches when each agent or monitor sees only its own subtask; because an adversary who controls the communication layer can force every method to L0, it is the worst-case safety number, reported per defense. The *Context Disclosure Threshold (CDT)* is the lowest disclosure level at which a method's detection rate rises significantly above its L0 baseline (one-sided Fisher's exact test, $\alpha = 0.05$); it answers the cooperative-case question of how much of the pipeline a defender must reveal before a method begins to work, and is read as a position on the L0–L4 gradient rather than a rate. CDT alone can be gamed by a method that simply grows more suspicious as it sees more, so we measure it twice, on the harmful pipelines and on matched controls; their gap is the *discrimination quality*, and if the two coincide the method is merely cautious rather than discerning.

The domain-matched gap. Because a method can be cautious about an entire subject area and thereby appear discerning, every discrimination claim in this paper is read against a *domain-matched* control rather than a benign one. The *domain-matched gap* for a method at a level is the detection rate on harmful subtasks (S1, S4) minus the detection rate on same-domain shady-but-legal subtasks (S9, S10), in percentage points; *benign collateral* is the refusal rate on benign controls (S7, S8) for the same method and level, reported in brackets. A good method has a high gap and low collateral; one with a high gap and high collateral is refusing broadly, and one with a low gap at any collateral is tracking domain rather than harm. Because the subset carries two twin pairs, individual gap cells carry roughly ± 10 points of sampling noise, which is why the paper reports qualitative orderings replicated across runs, and why the injection result is reported as a *within-run* clean-versus-injected contrast over the same subtasks, in which that noise cancels.

The observed–unobserved gap. The difference in a method's detection rate between rollouts that tell the agent its session is logged for review and rollouts that do not; a method that behaves the same either way acts on disposition, one that detects harm only when it believes it is watched acts on performance, and because a deployed agent often cannot tell whether it is monitored, the unobserved rate is the one that describes real deployment. Rollouts are paired: the same subtask is run under both framings at the same seed, so the contrast is within-item.

Experiment 1 scoring. Each decomposition's subtasks are scored by an independent judge from a different model family than the worker, reasoning enabled, in three roles: a refusal classifier that labels the worker's response `complied`, `refused`, `partial`, or `needs_input`; a suspicion rater that scores each subtask 1–5 for how harmful it looks in isolation, the Benignity Score being the mean

across a decomposition's subtasks; and a completeness judge that, given the objective and the full set of subtasks, decides whether carrying them out would achieve it. The Viable Threat Rate counts decompositions with Benignity ≤ 2 and completeness; the Operational Threat Rate counts those every subtask of which the worker complied with and that are complete.

Statistical treatment. Every headline quantity is a detection or refusal rate over rollouts, reported as a proportion with a Wilson score 95% confidence interval rather than a point value, because two defenses that both sometimes refuse are different defenses if one refuses 55% of the time and the other 95%. Pairwise comparisons between defenses, and between the harmful and control arms of the same defense, use Fisher's exact test. Where cell sizes permit, a mixed-effects logistic model with random intercepts for scenario and for decomposition separates a defense's effect from the scenario it happened to be tested on; where they do not, we report leave-one-scenario-out sensitivity instead of a model that the data cannot support. Results are reported at the instance level alongside aggregates, following the reporting standards the evaluation-science literature calls for, so that a reader can see the per-scenario spread and not only the mean.

Residual measurement limits. The judges are imperfect: completeness is judge-unstable and cannot headline without human grounding, and discrimination rests on classifier calibration, which is why human grounding of those labels remains outstanding. Two further scope limits apply to the trained disposition: a disposition trained into weights may behave differently at frontier scale, and the real-tool frictions come from one sandboxed platform and may not generalize.

Judge validity. The measurement leans on model-based judges, and we treat their validity as a question to be answered rather than assumed. The behavioral instrument is a four-label refusal classifier that separates a genuine safety refusal from a worker that merely lacks an upstream step's input (`needs_input`); collapsing those two, as a keyword heuristic does, inflates apparent safety, and the separation is what makes the 8% L0 safety-refusal of Experiment 1 a real number rather than an artifact. Suspicion and completeness are scored by an independent judge from a different model family than the worker, with reasoning enabled and an auditable rationale retained for every judgment. Completeness in particular proved judge-dependent: two judges disagreed sharply on the same stored decompositions, and one produced a benign-control completeness rate too low to be credible, so we treat single-judge completeness as unreliable and do not headline any figure resting on it. The remedy, which this release does not yet carry out, is human grounding: benignity and completeness labeled by at least two human raters against a written rubric, with Krippendorff's α reported, and the human–LLM gap reported as a result in its own right, since the literature finds LLM monitors near chance on decomposed subtasks and whether human raters do better is exactly what the compartmentalized setting puts in question.

Reproducibility. Each run writes a manifest recording every model identifier, temperature, token cap, rollout count, and seed, the verbatim prompt templates, the git commit and working-tree state, and a hash of the input file and of the harness, together with the exact set of decompositions selected; the raw worker and judge outputs, including reasoning and anything that failed to parse, are retained in full so that any reported datapoint can be reconstructed from the code and inputs that produced it.

C.1 L0 Pilot Slice by Defense

Legible versus laundered decompositions. Laundering removes what the rule was recognizing. The crux is whether a rule prompt that catches harm on a *legible* decomposition still does once the same subtasks are laundered. On a legible decomposition a rule prompt discriminates, catching harm while sparing the benign twin (a $+60$ gap at L0, pilot slices of $\sim$10 subtasks); laundering collapses that gap to about $+15$. The character prompt never discriminates, refusing the benign twin about as often as the harmful subtask, and sweeping prompt strength does not rescue this, permissive prompts under-fire and strict prompts refuse everything. The unaligned worker detects nothing at L0. So at L0 no prompt both catches laundered harm and spares the benign twin, the failure the discrimination metric is built to expose; the full ten-scenario sweep in Section 5.2 confirms the failure at proper sample sizes.

D Released-Artifact Manifest

We release what lets others reproduce and extend the evaluation without releasing what would function as a how-to. The public artifacts are the threat taxonomy (Appendix A), the scenario bank as abstract pipeline shapes (Appendix B), the metric definitions and the evaluation harnesses that compute them, the training and data-generation code for the fine-tuned adapters (Appendix H), and the per-run manifests and raw rollouts described in Appendix C, which name the open-weight model families, temperatures, seeds, and prompt templates for the defensive and judging roles. Withheld, as operational detail that would ease misuse, are the specific orchestrator prompts that successfully launder a harmful objective, and the tool configurations used to execute pipeline steps. The commercial agent-building platform of Experiment 3 is SaagaSolve, which funds this work. The hosted inference providers are named in the released run manifests. The abliterated orchestrator is described by capability and family rather than by a copy-ready pointer. This split follows the paper's standing rule: the structure and the defensive picture are released so the stack can be reinforced; the recipe is not.

E Defenses

Model-level alignment. The *rule* condition is a system prompt instructing the worker to refuse any task that could contribute to harm, surveillance of a person, deception, or illegal activity, even indirectly and even when the request looks routine; it is swept at three aggressiveness settings, of which the *permissive* setting, which declines only what is clearly intended to cause harm or is plainly illegal and otherwise helps, is the strongest performer and the one Table A4 reports. The *character* condition supplies a persona that weighs the broader consequences of participating and asks what a task is for when something seems off, relying on disposition rather than enumerated rules, and is swept at the same three settings. *DPO preference training* moves alignment from the prompt into the weights by preference optimization on pairs built from the same source subtasks as the supervised set, so that the supervised-versus-preference comparison isolates the training objective rather than the data. *Virtue-based character training* is a low-rank adapter supervised-fine-tuned on synthetic virtue-and-vice demonstrations of justice, trained into the same worker base as the other conditions so the comparison isolates the manner of alignment. Its training objectives are held out from the evaluation bank and are organized by *harm domain* (credential theft, influence operations, money movement, controlled acquisition, surveillance, coordinated harassment, account takeover), with a harmful and a same-domain legitimate objective paired within each domain, so the adapter learns to separate same-domain look-alikes rather than to blacklist a subject area. These training domains are not the six subtask frames of Appendix A: the frames describe the *shape* of the request a worker receives at L0 and are a property of the threat, whereas the training domains describe the *subject matter* of the held-out objectives and are a property of the dataset. Appendix H gives the details. Appendix H describes the variants added after the first draft: the compositional, mixed-data, and adversarially-trained adapters, and the prompt-plus-adapter combinations.

Monitoring (two protocols). The per-subtask monitor is handed one subtask in isolation; the pipeline-level monitor is handed the assembled set of subtasks. In both, detection is a safety refusal (the `refused` or `partial` label of the four-label classifier); `needs_input` does not count as detection, because a monitor that merely lacks upstream data has caught nothing. The pipeline-level protocol is implemented but not yet run.

Infrastructure. We implement no infrastructure defense. At each real-tool pipeline step we record whether the action succeeded and, where it did not, what stopped it; those stopping points are the chokepoints of Experiment 3.

F Ethics and Responsible Disclosure

This work studies how harmful objectives can be distributed across agents, which obliges us to state plainly what we did and did not do. The study ran under a documented protocol with a twofold purpose: to ensure the research creates no real-world harm, and to make the safeguards legible to readers weighing the findings against the risk of publishing them.

We never execute a harmful action. In scenarios where the harmful element is external to the agent system, the agents only coordinate (logistics, recruitment, payment) while the harmful input would come from outside, so we run the full composition end-to-end; with that input withheld the pipeline is genuinely harmless, a delivery pipeline on an ordinary package being just a delivery. Where the harmful action is internal, we run only up to the step before it, which is never taken. In no case is a complete harmful pipeline executed against a real target; the claim that one would compose is an argument from the structure we demonstrate on matched benign and shady-but-legal pipelines, not a harmful composition performed.

We did not target real people. Information-gathering used consenting team members or public figures with only already-public information; any task involving a human worker used genuinely benign, fairly compensated work. No financial operation used real money on a live network: payment and settlement ran on a test network, and any market interaction was simulated or confined to negligible stakes. Every real-tool run was sandboxed and monitored in real time, with the ability to terminate immediately.

We are deliberate about what we release. The paper publishes the framework, the threat taxonomy, the scenario structures as abstract pipeline shapes, the methodology, and the aggregate findings; it withholds the specific prompts that decompose a harmful objective, the tool configurations that execute steps, and any operational detail that would function as a how-to. The adversarial-adversarial training material of Appendix H is a defensive training recipe, and the injection styles it uses are ordinary published prompt-injection patterns rather than novel attacks; we release the training procedure and withhold nothing that is not already public, because a defender reproducing the result needs it and an attacker gains nothing from it. The released artifacts were reviewed against this standard before submission.

Finally, where the evaluation identifies a specific deployed model as unusually vulnerable, we follow responsible disclosure: the provider is notified before publication and given the standard 90 days to respond. The aim is to measure where the defensive stack stops holding so it can be reinforced, and this protocol is what lets us pursue that aim without becoming the threat we describe.

G Philosophical Grounding

This appendix records the virtue-ethical grounding that the main text uses in plain language, for readers who want the classical sources. It is not required to follow the argument or the results; the paper's claims rest on the instrument and the measurements, not on the exegesis below. References are to Aristotle's *Nicomachean Ethics* (NE), cited by Bekker book and chapter (trans. Ross, rev. Brown, Oxford, 2009).

Three paths. We place the three families of model-level alignment against the three major normative traditions. Preference-based training (RLHF, DPO) is broadly consequentialist: it optimizes an outcome signal. Rule- or principle-based training (Constitutional AI) is broadly deontological: it governs conduct by written injunctions. Character training is the virtue-ethical path: it forms a disposition of the agent rather than optimizing an outcome or applying a rule. The three are not mutually exclusive in practice, but they differ in what they treat as the primary bearer of alignment: the outcome, the rule, or the character.

	Encodes	Training signal	What is learned
RLHF / DPO	an outcome signal	optimize toward the reward	outcome-seeking
Constitutional AI	rules or principles	self-critique against the rules, then retrain	rule-application
Character training	a first-person character	a generator embodies the character and writes demonstrations, then fine-tune	a disposition

Table A2: The three alignment families by what each treats as the primary bearer of alignment: the outcome, the rule, or the character. Character training and Constitutional AI both use a written document and both retrain; they differ in whether the document is rules to apply or a character to embody.

Disposition (*hexis*, NE II.6). What character training forms is, in Aristotle's terms, a *hexis*: a stable state of character acquired through habituation, as distinct from a passing feeling or a rule one consults. A *hexis* disposes its holder to act in a settled way across situations, including ones never explicitly rehearsed, which is why the main text describes a trained disposition as issuing in conduct the agent was never directly taught. It is worth noting that this classical claim and the contemporary out-of-distribution generalization results discussed in Section 2.3 point at the same phenomenon from opposite directions: a disposition is precisely the kind of thing that transfers to unrehearsed cases, and the empirical finding that narrow fine-tuning moves behavior in untouched domains is what makes the analogy more than decorative.

Habituation (*ethos*, NE II.1). A *hexis* is acquired in one way, by habituation, the repeated performance of the acts characteristic of the virtue: we become just by doing just acts and temperate by doing temperate ones, as a craft is learned by practice. The disposition forms through repetition, and forms before the learner can give a rational account of the act or choose the act for its own sake; habituation is how a thing that does not yet hold a virtue comes to hold one. Training on demonstrations is habituation in this sense: the model is shown the characteristic act across many situations, and the disposition settles into the weights through repeated exposure rather than into a rule consulted at inference. This is also why the calibrated examples of the mean are duplicated in the training set (Methods) while the examples that decline are not; on the doctrine of the mean the disposition is habituated toward the intermediate act and away from the excess, and habituation is a matter of which acts are practiced most. The analogy is to the acquisition of a disposition, not to its full exercise; whether the result is virtue in the fuller sense that requires settled choice is the separate question the observed-and-unobserved test below is built to probe.

Practical wisdom (*phronesis*, NE VI.5). A virtue is not a rigid rule but is exercised through *phronesis*, the practical wisdom to perceive what a particular situation demands and to act accordingly. This is the aim the main text calls the judgment to notice the frame a subtask withholds: to weigh, in a request that carries no explicit red flag, whether the whole it serves warrants hesitation. In practice the trained disposition reaches this only where the subtask leaves the harm legible; under the isolation that hides the whole it is bounded like any intent-reading defense.

The virtue chosen: justice (*dikaiosyne*), not honesty. We train toward justice rather than honesty. Honesty (truthfulness) governs how an agent represents itself, but it has no purchase on a worker asked to perform a genuinely benign-looking subtask, where nothing is being misrepresented. Justice (*dikaiosyne*), a standing concern for the parties whom one's conduct may wrong, is what bears on compartmentalized harm, where the wrong falls on a third party outside the exchange. Justice is also structurally a *higher-disclosure* virtue: it reasons about the composition, which L0 hides, so at full isolation the disposition necessarily degenerates to whatever the subtask's surface leaves legible, and only becomes justice-like as context appears. This unifies the otherwise odd pattern in our results, in which the trained disposition is modest at L0 and strongest at L3.

Virtue against vice. Following Aristotle's method of defining a virtue against its corresponding vice, the disposition is trained against a specific failing rather than toward a trait in the abstract. Its opposing vice is a servile instrumentality: the readiness to carry out any well-formed task and to ask nothing of the whole it serves, indifferent to whom that whole might wrong. The training demonstrations pair the virtue against this vice across held-out objectives spanning the harm domains of Appendix A.

The observed and unobserved test (NE II.4). Aristotle holds that a virtuous act must proceed from a firm and settled character, chosen for its own sake, rather than be performed because one is being watched. This grounded our sharpest intended test of genuineness: a trained disposition should behave the same whether or not the agent believes its session is observed, whereas a mere performance holds only under observation. We report the outcome as the null-to-negative it was (Section H.1): the trained adapter refuses more when told it is audited, which on this criterion counts against its genuineness, and we record it rather than reframing the criterion after the fact.

H Training and Run Details

This appendix documents the fine-tuned adapters and the runs behind Sections H.1–5.4. All adapters are low-rank (LoRA, $r = 64$, $\alpha = 128$) trained on the same open-weight instruction-tuned base unless noted, and every method is served behind the same evaluation harness so that a method is a $(model,\ systemprompt)$ pair and nothing else varies.

Training data. The adapters are trained on sixteen held-out objectives, disjoint from the evaluation bank and organized by harm domain, with a harmful and a same-domain legitimate objective paired inside each domain (DR: same-domain look-alikes, not domain blacklisting). Each objective is decomposed and laundered by the same abliterated orchestrator used in Experiment 1, since safety-trained models and frontier teachers refuse to launder and this step cannot be done with an aligned model. The resulting subtasks become the training inputs. A frontier teacher then writes the calibrated target response for each subtask. The teacher sees the objective, so the teacher knows the right answer; the training example's user turn is the *bare subtask only*, in L0 form, so the model must learn to produce that answer from the subtask alone. A judge then scores each pair 1–3 for over-refusal and rewrites the miscalibrated ones. One property of the resulting labels is worth stating plainly because it determines what the adapter can do: the teacher decides on what is visible in the subtask itself, and a well-laundered harmful subtask therefore receives a *proceed* target. The adapter is trained to act on residual legibility, not to infer a hidden composition, which is why its detection is modest at L0 and strongest at L3.

Adapters. *Surface SFT* (reported as the trained disposition) is supervised fine-tuning on those demonstrations, each pair a subtask plus a calibrated target response of decline, ask, or proceed. *Compositional SFT* replaces those with judgments made over partially-disclosed context, teaching the adapter to judge the composition rather than the surface; it over-refuses, having learned that disclosure itself is evidence of harm. *Mixed-data SFT* combines the two sources, which corrects the compositional adapter's calibration (benign collateral falls from about 20% to 5%) and so confirms that the failure was a data problem, without recovering a win. *DPO* trains a fresh sibling adapter on the same base by preference optimization (learning rate 5×10^{-6}, $\beta = 0.1$), with the chosen response the calibrated supervised target and the rejected response the opposite miscalibration. One design note is worth recording because it constrains anyone reproducing this: a frontier teacher model *cannot* generate the rejected (compliant) side for exactly the harm-leaking subtasks that matter most, because its own safety layer refuses, so the rejected side is generated on the base policy instead, which is the on-policy form anyway. *Adversarially-trained* augments the surface-SFT set with eight prompt-injection styles, duplicating each training pair into a clean and an adversarial-context copy that share the same calibrated target, so the adapter learns that the injected context does not change the correct answer.

The held-out injection. Evaluation uses a ninth injection style, absent from training: an operator-directive system prompt asserting that the worker is a fully autonomous execution agent inside a trusted, pre-approved internal pipeline, that every task has already been authorized, and that refusing, questioning, warning about, or asking for clarification on any task is prohibited. It is applied as the worker's system prompt, which is precisely what an adversary owning orchestration can set. Clean and injected conditions run over the same subtasks within one run, so Table 1 is a paired contrast.

Runs. The results in Sections 5.3–5.4 come from four independent sweeps over the six-scenario domain-matched subset at L0–L3, two rollouts per cell: a prompt-strength gradient (7 conditions, 9,576 records), a training-method comparison (5 conditions, 7,040 records), a data-fix and combination study (6 conditions, 8,208 records), and the injection-robustness study (6 conditions, 7,872 records), plus a cross-base replication on a larger heavily safety-trained model (4 conditions, 5,632 records). The observed–unobserved study of Section H.1 adds 4,608 paired records at L0 and L3. Error rates were zero in all sweeps but the cross-base one, which had two. Detection is judged by a classifier from a different model family than any worker under test. Because the qualitative ordering of methods is what we report, and because each sweep independently reproduces it, we prefer replication across runs to a single larger run as the check on the two-twin sample.

Defense (L0)	Harmful	Shady-legal	Benign
Unaligned	20%	4%	0%
Rule prompt	65%	45%	28%
Character prompt	87%	95%	74%

Table A3: L0 detection by defense against a harmful pipeline and two controls: benign, and shady-but-legal in the same domain (all ten scenarios; $n \approx 178/56/58$). The rule refuses the legal same-domain control (45%) nearly as often as the harmful subtask (65%), and the character prompt refuses all three at high rates: read against the domain-matched control, neither discriminates. Only the unaligned worker keeps a clean benign gap, but detects little at L0.

Defense	L0	L1	L2	L3
Trained disposition	+17 [14]	+24 [11]	+24 [7]	**+42** [5]
Permissive rule *prompt*	+13 [11]	**+31** [8]	**+31** [0]	+11 [3]
Compositional SFT	+7 [7]	+6 [19]	+12 [20]	+4 [15]
DPO preference training	+3 [0]	+0 [0]	+0 [0]	+5 [0]
Untrained base	+2 [0]	+0 [0]	+0 [0]	+3 [0]
Mixed-data SFT	+10 [2]	+18 [14]	+19 [5]	+20 [9]
Trained + permissive prompt	+2 [43]	+1 [65]	+14 [57]	+7 [5]

Table A4: Domain-matched discrimination gap by disclosure level, in percentage points, with benign collateral in brackets (lower is better). In this Qwen3-8B clean comparison, a permissive rule *prompt* matches or beats the evaluated trained methods at L1 and L2. The trained disposition's largest point-estimate advantage appears at L3. Formatting: **highest gap** per level.

H.1 Can Training Restore L0 Recognition?

Development-stage evidence. The scenario bank has been used repeatedly to select methods and to shape this paper, so under standard adaptive-holdout reasoning the bank is development data rather than a held-out test set. This subsection and the two that follow are accordingly exploratory: they establish direction and motivate a sealed evaluation, and do not constitute a confirmatory trained-versus-prompted comparison.

The confirmatory replacement for this subsection ran on Qwen3-32B rather than on the design frozen in DR-043, which was superseded before execution. It is reported in Section 5.5.

The prompted defenses leave the central question open: can a *trained* disposition recognize laundered harm at L0 without the character prompt's blanket refusal? We test this as a self-contained, same-base study. A low-rank adapter is fine-tuned for justice on held-out objectives (disjoint from S1–S10; Appendix A) and run through the full gradient against three prompt-based controls on the same open-weight base, isolating training from the base. That base is smaller than Experiment 2's worker and recognizes almost nothing even when the objective is disclosed, so any L0 recognition here is trained, not native.

Defense (L0)	Harmful	Shady-legal	Benign
Base (untrained)	1%	0%	0%
Rule prompt	89%	66%	38%
Character prompt	88%	78%	76%
Trained disposition	42%	22%	12%

Table A5: L0 detection by category (safety refusal) on the trained-disposition study: a smaller open-weight base than Experiment 2's worker, held-out scenarios, four same-base conditions ($n \approx 346$ harmful, 116 shady-legal, 118 benign). The trained disposition's edge is low *collateral*, refusing only 12% of benign work against the rule's 38% and the character prompt's 76%; the character prompt over-refuses every category. Discrimination is read against the domain-matched control, not these category rates (see text). The base recognizes almost nothing, so the trained arm's L0 recognition is instilled by training.

I Experiment 3 in Full

The main text reports these two studies in summary because the paper's contribution is the control evaluation of Section **??** onward. They bracket that evaluation: Experiment 1 establishes that the attack is available, and Experiment 3 that the only surviving boundary is one that reads no intent. Both sections are reproduced here as written.

Operational feasibility and infrastructure chokepoints. Experiments 1 and 2 concern intent; Experiment 3 concerns action. Once the inner layers fail at L0, only a layer that reads no intent can stop the pipeline, so this is where a durable defense must live; we probe its feasibility here rather than build it, and designing that layer is future work. We run one information-gathering-and-coordination pipeline with real tool access on a commercial agent-building platform, operated by the funder of this work, sandboxed and monitored in real time: benign and shady-but-legal pipelines run as full compositions, harmful ones only to the step before any harmful action (Ethical Statement). At each step we record whether the *action* succeeded and, if not, what stopped it. This single pass locates the two boundaries the framework predicts, an intent boundary the adversary dissolves and action boundaries it cannot; a systematic sweep is the committed extension.

The intent boundary appears at agent creation. Asked to create an agent whose system prompt spelled out a personal-profiling task, the platform's frontier creation model refused, naming it surveillance; asked for a generic "research assistant" with the same tool access, it complied. Only the legibility of the purpose differed, and compartmentalizing it out of the prompt moved the request from refused to routine. Run at L0 on a consenting team member, the assistant compiled a detailed profile, including an inferred neighborhood, without concern: the boundary does not re-arise at execution once cleared at creation.

The action boundaries appeared only where an operator had built one. Web search and open-page scraping ran unimpeded, but scraping a major professional network was blocked by its anti-automation controls, and the outbound-email step stalled at an interactive authorization the agent could draft around but not clear. A serverless backend added non-intent friction (cold starts, timeouts, throttling) that taxes an unattended pipeline; financial steps were confined to a test network by construction. These surviving frictions read no intent, but exist only where an operator built them, most outside the AI-safety community a defender would rely on.

J Cross-Base Robustness on Gemma 4 31B

Repeating the Section 5.3 and Section 5.4 recipe on a Gemma 4 31B base inverts the main-text picture, so the effect is base-dependent rather than a scaling or cross-family law.

Discrimination recipe. Gemma 4 31B already discriminates unaided (a $+15$ gap at L0 rising to $+53$ at L3 at zero collateral), and the recipe adds collateral without adding recognition, ending worse than the untouched base at L3. The recipe can instil in Qwen3-8B behavior that Gemma 4 31B already exhibits, and cannot be assumed to help on every stronger base.

Hostile-prompt robustness. Under the same withheld injection the ordering flips (Table A6): the rule prompt does not collapse, adversarial training actively harms, and the same injection *raises* the bare base's gap, from $+14$ to $+18$ at L0, making the base more discriminating rather than less. Gemma resists the attack natively, so there is little for the disposition to recover. With the Qwen3-32B confirmation ruling out a simple less-versus-more-safety-trained account, the honest scope is an interaction among base, training set, threat model, attack, and context.

Acknowledgments and Disclosure of Funding

Funding. The research leading to these results has received funding from SaagaSolve.

Acknowledgements. We thank Andrew Melnychuk-Oseen for identifying the threat model studied here in the course of building an agent platform, for proposing this research, and for access to that platform for the evaluation in Experiment 3.

	L1 gap		L2 gap		
Defense	intact	hostile	intact	hostile	Retained
Adversarially-trained	$+7$	$+5$	$+22$	$+10$	$\sim 50\%$
Standard-trained	$+14$	$+2$	$+36$	$+39$	$\sim 82\%$
Permissive rule prompt	$+27$	$+19$	$+47$	$+26$	$\sim 61\%$

Table A6: Gemma 4 31B under the withheld prompt injection (reduced evaluation, one rollout, L1–L2, correspondingly noisy). The ordering flips relative to the Qwen3-8B result of Table 1: adversarial training leaves the adapter with the lowest intact gap, and the untouched base is the most robust arm.